ČASOPIS PRO MODERNÍ FILOLOGII 103, 2021, Č. 2, S. 309–314

OPEN ACCESS

LEMUR — DATABÁZE VÍCESLOVNÝCH LEXIKÁLNÍCH JEDNOTEK ZPRÁVA O ELEKTRONICKÉ DATABÁZI

Na podzim roku 2020 byl publikacemi Petkevič et al. (2020) a Rosen et al. (2020) oficiálně ukončen čtyřletý projekt *Mezi slovníkem a gramatikou*,[1] jehož cílem bylo vytvořit reprezentativní databázi klasifikovaných víceslovných jednotek v češtině (dále VLJ) o rozsahu nejméně 7000 hesel. V projektu tedy vznikla

(i) databáze jakožto softwarový systém
(ii) lingvistická typologie VLJ inspirovaná zahraniční a domácí literaturou a rozsáhlými korpusovými daty.

O aktuálním stavu databáze a typologii víceslovných jednotek referoval za celý tým (Milena Hnátková, Tomáš Jelínek, Alexandr Rosen, Hana Skoumalová a Vladimír Petkevič z Ústavu teoretické a komputační lingvistiky FF UK; Marie Kopřivová a Pavel Vondřička z Ústavu Českého národního korpusu FF UK) vedoucí projektu Vladimír Petkevič na úterním semináři Ústavu Českého národního korpusu FF UK 8. prosince 2020.[2]

Víceslovnými jednotkami nazývají tvůrci databáze (slovníku) LEMUR ustálená významově jednotná spojení, která se skládají nejméně ze dvou slov psaných zvlášť. Hrají v jazyce významnou roli a mnoho z nich se v úzu vyskytuje nadprůměrně často. V jazykovém systému jsou to jednotky nacházející se na pomezí slovníku a gramatiky. V literatuře se označují různě (víceslovná pojmenování, víceslovné lexémy, slovní spojení, sousloví). Do databáze jsou zahrnuty také statisticky významné kolokace (srov. níže frekvenční/empirický přístup).

K březnu roku 2021 obsahuje databáze VLJ okolo 10 200 hesel a stále se rozšiřuje, doplňuje a upřesňuje. Jednotlivá hesla jsou popisována a klasifikována na základě podrobné typologie VLJ, jež zachycuje jejich vlastnosti na těchto jazykových rovinách/plánech: morfologie, syntax, sémantika, pragmatika, lexikon. Každá VLJ je charakterizována bohatým repertoárem vlastností s cílem:

— VLJ komplexně klasifikovat z hlediska jejího postavení v jazykovém systému češtiny; klasifikace je užitečná mj. pro zkvalitnění morfologické anotace a syntaktické analýzy jazykových korpusů, disambiguaci lexikálních významů, sémantické značkování, lexikografii a lexikologii;
— v budoucnu propojit databázi s korpusy a umožnit tak anotovat korpusy víceslovnými jednotkami z databáze;
— na základě anotace umožnit, aby se VLJ dala v korpusu rozpoznat a studovat nejen ve své standardní, kanonické podobě, ale i v podobě variant a fragmentů různého typu (morfologických, syntaktických, lexikálních), a to na základě dotazů na různé vlastnosti VLJ a/nebo jejich kombinace.

V databázi LEMUR se autoři pokusili spojit dva přístupy (srov. Klégr, 2016):

1 Podpořila jej Grantová agentura České republiky, reg. č. 16-07473S.
2 https://trnka.ff.cuni.cz/seminar.

ISSN 0008-7386

— *frazeologický/sémantický*: VLJ jsou ustálené syntagmatické konstrukce vymezitelné svými charakteristickými vlastnostmi (top-down approach);
— *frekvenční/empirický*: VLJ se vymezují podle frekvence výskytu a jeho statistické významnosti jako opakující se souvýskyty slov zjišťované analýzou korpusových dat (bottom-up approach).

Výzkumný tým přitom vycházel z těchto datových zdrojů:
— Slovník české frazeologie a idiomatiky (Čermák et al., 1983–2009) — pro frazémy;
— FRANTALEX — seznam frazémů a kolokací (Milena Hnátková);
— korpusy současné češtiny řady SYN v Českém národním korpusu.

Zvolenou typologii inspirovala klasifikace VLJ navržená v článku Baldwin et al. (2010) a autoři rovněž vycházeli z návrhu projektu PARSEME (http://typo.uni-konstanz.de/parseme), který na tento článek navazuje. VLJ jsou v něm kategorizovány podle tří hlavních kritérií:

a) syntaktická struktura
b) ustrnulost/flexibilita
c) idiomatičnost[3] (včetně různých nepravidelností ve VLJ).

Autoři LEMURu toto třídění přijali a rozšířili se zřetelem

(i) k osobitým vlastnostem češtiny:
   — morfologická idiomatičnost (daná velmi bohatou a spletitou morfologií češtiny)
   — syntaktická specifika (zvláště volný slovosled)
(ii) k dvojí povaze databázových hesel:
   — jsou užitečná pro lidského uživatele
   — lze jich též využít v počítačových aplikacích.

Rozšíření zahrnuje tyto údaje a vlastnosti: definici, charakteristické příklady, typ užití, fragmenty a varianty VLJ, styl/varietu VLJ a rovněž podrobnější charakteristiku některých vlastností, zvláště těch, jež jsou typické pro češtinu.

Velmi ambiciózní je zejména snaha zachytit fragmenty a varianty standardních VLJ. Při stanovení invariantu frazému a jeho variant odvozených od invariantu vycházejí autoři z jádra VLJ, které chápou buď jako kombinaci minimálně dvou lexikálních prvků, z nichž jeden může být proměnlivý, anebo je to struktura, jejíž lexikální obsazení podléhá určitým omezením. Máme-li například standardní biblickou VLJ: *Snáze projde velbloud uchem jehly než bohatý do Božího království*, chápe se trojice slov *velbloud uchem jehly* jako její fragment a její modifikovaná podoba v korpusu *To by spíš velbloud vešel do království nebeského, než abych já navlékl nit uchem jehly* jako její

3 V databázi chápou autoři LEMURu idiomaticitu jako „obecnou vlastnost" frazému, která se vztahuje k různým jazykovým rovinám (lexikální, morfologické, syntaktické, pragmatické) a zřetelům (frekvence).

aktualizující varianta. Kromě takovýchto fragmentů a variant hodlají autoři též zachytit takové VLJ, jež vznikly syntaktickými transformacemi ((de)pasivizace, (de) substantivizace, (de)adjektivizace) nebo syntaktickými modifikacemi ((ne)možností syntakticky rozvíjet určité slovo ve VLJ).

V charakteristice VLJ se obecně předpokládá platnost standardních gramatických pravidel češtiny a zachycují se pouze odchylky od těchto pravidel.

Heslo VLJ v databázi LEMUR obsahuje tyto údaje (podrobnosti viz Petkevič et al., 2020):

— *lemma* jednoznačně identifikující VLJ jak v databázi, tak v anotovaném korpusovém textu, například *dostat_přes_kušnu*;
— *superlemma* jednoznačně sdružující skupinu nějakým způsobem příbuzných VLJ, identifikovaných svými lemmaty, např. *čest_a_sláva* (do příslušné skupiny patří třeba VLJ *umírat pro čest a slávu*, *bojovat za čest a slávu*…);
— *lemmata* a *morfologické vlastnosti* jednotlivých komponent (slov) VLJ;
— *definice* objasňující význam VLJ;
— *styl/varieta* popisující VLJ podle stylu/registru: spisovný / kolokviální / nářeční / expresivní / slang / jiný;
— *typ užití* charakterizující VLJ z tradičního frazeologického hlediska: přísloví / pranostika / přirovnání / citace / cizojazyčné spojení / termín / víceslovné synsémantikum / nespecifický slovesný frazém / neslovesný frazém / kvazifrazém / větný frazém / otevřený frazém / uzuální kolokace;
— *syntaktická struktura*, vystižená těmito údaji:
  ▪ *syntaktický typ* kategorizující VLJ podle syntaktické kategorie celé jednotky: jmenná skupina / adjektivní skupina / plnovýznamová slovesná fráze / slovesná fráze s kategoriálním slovesem / adverbiální skupina / předložková skupina / složená předložka / složená spojka / složené citoslovce / klauze / souvětí / jiný složený výraz;
  ▪ *základní strukturní vzorec* jako posloupnost kódovaných slovních druhů vyvozená ze syntaktického stromu VLJ;
  ▪ *syntaktický strom*: závislostní i frázový, obsahující i příslušné syntaktické funkce;
  ▪ *adverbiální sémantický typ*: místní určení / časové určení / určení způsobu / okolnostní určení;
— *ustálenost/flexibilita* zachycená těmito údaji:
  ▪ lexikální a další varianty VLJ; VLJ se totiž zdaleka nemusí vyskytovat pouze ve své standardní podobě;
  ▪ relevantní fragmenty standardní VLJ;
  ▪ slovosledná specifika;
  ▪ vnitřní modifikovatelnost: možnost syntakticky rozvíjet některou z komponent VLJ;
  ▪ syntaktické transformace: (ne)možnost nominalizace, (ne)možnost adjektivizace, (ne)možnost pasivizace či aktivní podoby;
  ▪ morfologická omezení: morfologické zvláštnosti jednotlivých komponent (slov) VLJ dané právě jejich přítomností ve VLJ;

— *idiomatičnost* dále členěná na idiomatičnost
  - *lexikální*: slova monokolokabilní / slova téměř monokolokabilní / negativa tantum / cizojazyčné výpůjčky / makaronismy / jiná;
  - *morfologickou*: morfologicky nestandardní tvary vyskytující se pouze v příslušné VLJ;
  - *syntaktickou*: anakolut / atrakce / aposiopeze / elipsa / zvláštní valence / zvláštní slovosled / jiná;
  - *sémantickou*, odrážející míru významové kompozicionálnosti složek VLJ: plně kompozicionální / často kompozicionální / zřídka kompozicionální / nekompozicionální;
  - *pragmatickou*, charakterizující užití VLJ ve specifické situaci;
  - *statistickou*: nadprůměrně frekventovaná kolokace s nápadně omezenou kolokabilitou.

Softwarová koncepce lexikální databáze, kterou naprogramoval Pavel Vondřička, je dána především snahou zachytit variabilitu VLJ podle různých stupňů specifičnosti, např. variabilitu jednoho či více slovních tvarů konkrétního lexému, nebo variantní lexémy či různé tvary odlišných lexémů nebo frází různých typů apod. Návrh databáze počítal s tím, že je nutno zachytit jak vlastnosti celé VLJ (např. syntaktický typ celé VLJ), tak vlastnosti jednotlivých komponent VLJ (např. morfologická idiomaticita je vlastností jedné komponenty VLJ); navíc například styl může záviset na konkrétní komponentě nebo být vlastností celé VLJ (např. kolokviální může být celá VLJ, ač obsahuje jen spisovná slova). Databáze je navíc koncipována se zřetelem ke dvěma cílům: lexikografickému (pro lidského uživatele) a technickému (může například sloužit k automatické syntaktické analýze).

Variabilitu zachycují tři konstitutivní prvky ve struktuře lexikálního hesla:

— *sloty* — jednotlivé komponenty VLJ (syntagmatická dimenze);
— *náplně* zachycující možné typy, jež mohou být variantami komponenty; daný slot tak může být zaplněn několika různými náplněmi (dimenze paradigmatická);
— *vlastnosti/rysy* — dvojice typu name=value, která se dá přiřadit heslu VLJ nebo konkrétnímu slotu či náplni.

Slot je tedy výčet možných náplní, přičemž může být trojího druhu:

— *pevný/fixed* — komponentu lze realizovat jen jedním z vyjmenovaných typů;
— *otevřený/open* — komponentou může být libovolný lexém (v konkrétním tvaru);
— *zčásti otevřený/semi-open* — komponenta se obvykle realizuje jedním z lexémů v seznamu, ale někdy se mohou objevit synonyma nebo sémanticky příbuzné výrazy.

V databázovém hesle jsou zachyceny dva typy komponent:

— *jeden token* zachycený slotem s jednou či více náplněmi, přičemž náplň definuje určitý typ pomocí pozičních atributů (lemma, tag); tag je vyjádřen jako prefix morfologicky značkovaného slova, např.

- určitý pád, singulár nebo plurál: lemma="ryba", tag="NNF[SP]1";
- jakýkoli slovesný tvar: lemma="stát", tag="V";
- jakékoli obecné adjektivum: tag="AA";

Náplně tak reprezentují konkrétní typy, které rozpoznává syntaktický analyzátor;

— *fráze* jsou zachyceny otevřenými sloty bez náplní; vlastnosti přiřazené slotu definují typ a morfologická omezení kladená na obsah slotu. Sloty tak mohou reprezentovat abstraktnější jednotky pro účely lexikografického popisu (pro lidské uživatele) nebo pokročilého automatického parsingu. Dokážou rovněž zachytit variabilní komponenty tvořené několika tokeny (např. variabilitu reflexivního slovesa/adjektiva/substantiva a jeho nereflexivního protějšku) a stromové struktury závislostní i frázové (sloty neterminální).

V databázovém hesle lze také prostřednictvím odkazu zachytit opakování téhož (variabilního) lexému, například ve VLJ *Bůh dal, Bůh vzal*, kde *Bůh* může být pouhou instancí proměnné X ve struktuře *X dal, X vzal*; dají se tak efektivně zachycovat typy VLJ vykazující danou strukturu.

Vývoj databáze rozhodně není ukončeným projektem uzavřen, neboť

— databáze se stále rozšiřuje jak *extenzivně*: doplňují se nová hesla z hlediska frazeologického i frekvenčního, tak *interně*: zjemňuje/doplňuje/opravuje se popis již existujících neúplných hesel;
— se zjemňuje či může zjemňovat sama lingvistická typologie, a to *rozšiřováním* repertoáru hodnot určité vlastnosti a/nebo *doplňováním* nových vlastností (např. původ přísloví/rčení, cizojazyčné ekvivalenty…);
— v nejbližší době autorský tým plánuje propojit databázi s korpusy současné češtiny, což by umožňovalo využít potenciálu podrobného značkování databázových hesel a jejich (variantních) projevů pro vyhledávání v korpusových textech.

V budoucnu se autorský tým hodlá mimoto zaměřit na:

— vyhledávání dalších VLJ v korpusových datech, anotaci těchto VLJ a jejich zařazování do databáze LEMUR
— zpřesňování představené typologie
— publikaci dalších poznatků o vlastnostech VLJ, a to zvláště na datovém základě korpusů SYNv8, SYNv9.

LITERATURA

Baldwin, T. — Kim, S. N. (2010): Multiword Expressions. In: N. Indurkhya — F. J. Damerau (eds.), *Handbook of Natural Language Processing*, 2. vyd. Boca Raton: CRC Press, s. 267–292.

Čermák, F. et al. (1983–2009). *Slovník české frazeologie a idiomatiky* (*SČFI*), vol. 1–4. Praha: Academia/Leda.

Klégr, A. (2016): Lexikální kolokace: základní přehled o vývoji pojetí.

OPEN ACCESS

Časopis pro moderní filologii, 98, 1, s. 95–103.

Petkevič, V. — Kopřivová, M. — Hnátková, M. — Jelínek, T. — Kopřiva, P. — Rosen, A. — Skoumalová, H. — Vondřička, P. (2020): Typologie víceslovných jednotek v češtině a frekvenční zastoupení jejich hlavních vlastností v žánrově vyváženém korpusu. *Studie z aplikované lingvistiky*, 2, s. 37–62.

Rosen, A. — Skoumalová, H. — Znamenáček, J. (2020): Víceslovné lexémy v syntaktickém kontextu. *Studie z aplikované lingvistiky*, 2, s. 63–84.

Vladimír Petkevič | Ústav teoretické a komputační lingvistiky, Filozofická fakulta Univerzity Karlovy | Celetná 13, 110 00 Praha 1
ORCID ID: 0000-0003-0468-4158
vladimir.petkevic@ff.cuni.cz